Article:
Thelwall, M. orcid.org/0000-0001-6065-205X and Holyst, J.A. (2023) Can journal reviewers dependably assess rigour, significance, and originality in theoretical papers? Evidence from physics. Research Evaluation, 32 (2). pp. 526-542. ISSN 0958-2029

https://doi.org/10.1093/reseval/rvad018

Can journal reviewers dependably assess rigour, significance, and originality in theoretical papers? Evidence from physics[1]

Mike Thelwall, University of Sheffield; Janusz A. Hołyst, Warsaw University of Technology.

Peer review is a key gatekeeper for academic journals, attempting to block inadequate submissions or correcting them to a publishable standard, as well as improving those that are already satisfactory. The three key aspects of research quality are rigour, significance, and originality but no prior study has assessed whether journal reviewers are ever able to judge these effectively. In response, this article compares reviewer scores for these aspects for theoretical articles in the SciPost Physics journal. It also compares them with Italian research assessment exercise physics reviewer agreement scores. SciPost Physics theoretical articles give a nearly ideal case: a theoretical aspect of a mature science, for which suitable reviewers might comprehend the entire paper. Nevertheless, intraclass correlations between the first two reviewers for the three core quality scores were similar and moderate, 0.36 (originality), 0.39 (significance), and 0.40 (rigour), so there is no aspect that different reviewers are consistent about. Differences tended to be small, with 86% of scores agreeing or differing by 1 on a 6-point scale. Individual reviewers were most likely to give similar scores for significance and originality (Spearman 0.63), and least likely to for originality and validity (Spearman 0.38). Whilst a lack of norm referencing is probably the biggest reason for differences between reviewers, others include differing background knowledge, understanding, and beliefs about valid assumptions**.** The moderate agreement between reviewers on the core aspects of scientific quality, including rigour, in a nearly ideal case is concerning for the security of the wider academic record.

Keywords: Academic peer review; Journal quality control; Journal article referees; Peer review; Scholarly communication.

Introduction

Peer review became institutionalised over the past half century as a cornerstone of academic research, with journal articles, papers in serious conferences, book chapters and monographs typically needing to satisfy multiple reviewers and an editor before publication. Whilst expert reviews can help authors improve their work (Chong and Mason, 2021; Garcia-Costa et al., 2022), their primary role is often to judge the quality of submissions to ensure that only valid and (normally) useful work enters the scholarly record. Despite these critical functions of peer review, experts are usually untrained in reviewing (e.g., Freda et al., 2009; Warne, 2016), and routinely disagree substantially in their judgements (see Background), leaving authors and editors with conflicting recommendations (Peterson, 2020). It is therefore unsurprising that most of the 1,340 respondents to a survey of biomedical researchers considered journal peer review to be unscientific (Ho et al., 2013), suggesting that the scholarly record is unreliable. Exacerbating this problem, the prevalence of closed peer review (Wolfram, et al., 2020) keeps most disagreements hidden, creating a false impression of certainty. One study found that 8

[1] Thelwall, M. & Holyst, J. (in press). Can journal reviewers dependably assess rigour, significance, and originality in theoretical papers? Evidence from physics. *Research Evaluation*. https://doi.org/10.1093/reseval/rvad018

out of 9 articles reviewed by a journal were rejected, five with serious validity concerns, despite having already been published in the same journal (Peters and Ceci, 1982). This shows that reviewer discrepancies can translate into flawed editorial decisions. Even with these fundamental problems, the reasons why reviewers disagree are poorly understood (Tennant and Ross-Hellauer, 2020), undermining attempts to improve the current system.

All academic judgements are subjective, with implicit assumptions (Strevens, 2021), so there are many reasons why reviewers may disagree when evaluating academic research. These include variations in the amount of time spent reviewing, reviewer knowledge areas and levels, interpretation/ignoring of the reviewing guidelines, and beliefs about what is important in research and the role of reviewing. This is complicated by article quality being multidimensional, with reviewers potentially overlooking some dimensions. The most important aspects of quality are usually believed to be rigour, originality, and significance, with significance sometimes split into societal and academic components (Bonaccorsi, 2018, p. 82; Langfeldt et al., 2020). Of these, rigour is arguably the most critical and least subjective, so the scholarly record may be relatively safe from errors if inter-reviewer disagreement is primarily confined to originality and significance, or other facets (e.g., grammar). Nevertheless, with one partial exception (Oxman et al., 1991) no previous study has investigated the extent to which reviewers agree on each of the three dimensions of quality. Evidence about this is therefore needed to assess the seriousness of reviewer disagreements.

This article investigates for the first time how much reviewers agree on the three core dimensions of quality for one narrow type of research: theoretical articles in the journal *SciPost Physics*. The SciPost Physics online journal practices open peer review (with optionally anonymous reviewers) and includes reviewer estimates of six quality dimensions. The SciPost website appears to be currently the only public source of quality dimension ratings for journal article submissions. Fortunately, theoretical physics is a good topic for evaluating inter-reviewer consistency because reviewers seem more likely to be able to understand all aspects of submissions than is typical for science. This is because theoretical physics is mathematical and formulae and the assumptions behind them should be relatively understandable to the expert, even if unaware of the underpinning theory. This contrasts with, for example, cell biology research having large authorship teams contributing differing specialist results (e.g., Western Blots, genetic information, cell characteristics). In addition, physics is a relatively mature science, which might therefore have an above average consensus, with less scope for ideological conflicts between reviewers (Shepherd and Challenger, 2013; Whitley, 2000). Of course, there are still some disputes in theoretical physics, such as for string theory (Ritson, 2021). The results are also compared to apparently the only public dataset of reviewer scores for a national research evaluation exercise, a sample from the Italian Valutazione della Qualità della Ricerca (VQR) to contextualise the overall level of agreement. The following questions drive the study.

- RQ1: How consistent are the quality facet scores from different reviewers for SciPost Physics theoretical articles?
- RQ2: How does inter-reviewer consistency for SciPost Physics theoretical articles compare to inter-reviewer consistency scores from the Italian national research evaluation for physics articles (the only public comparable set of data from expert judges)?
- RQ3: Which quality facet scores attract the most similar scores from the same reviewer for SciPost Physics theoretical articles?

- RQ4: Why do reviewers disagree about the different quality components for SciPost articles?

Background

This section discusses how research quality is conceived by academics and the extent to which experts agree on it. Since there are substantial disciplinary differences in research objects, objectives, and practices, both issues are complex. The focus here is on the gatekeeping role of journal reviewing, although grant reviewing may have larger problems with reviewer consistency (Jerrim and Vries, 2020). There are relatively few theoretical contributions to peer review research (Hug, 2022), so this section primarily focuses on empirical findings. It analyses review without deliberation between reviewers, although some grants and computing conferences and national research evaluation exercises encourage reviewer discussions to resolve differences.

Dimensions of research quality and the role of peer review

Research quality is important for scholars, knowledge communities, academic institutions, funders, and policy makers. It is operationalised differently between fields and has differing connotations for policy and academia (Langfeldt et al., 2020). The general term "quality" allows the meaning of "research quality" to vary between academic contexts. This fluidity of meaning is exacerbated by quality judgements frequently drawing upon tacit knowledge applied to unique research objects, although advice is sometimes provided in formal contexts, such as reviewer guides for journals (Seeber, 2020) or quality criteria for national research evaluation exercise (e.g., Bonaccorsi, 2018, p. 82).

Examinations of journal reviewer guidelines in a few fields have identified over a thousand items that reviewers have been asked to consider (Capaccioni and Spina, 2018; Davis et al., 2018; Maggin et al., 2013; Song et al., 2021), and these may be thought of as the intricate components of research quality in some contexts. The Equator Network (www.equator-network.org) also provides reporting guidelines or checklists for a wide range of methods used in health research and this presumably informs authors, reviewers, and journal editors. For example, a web survey article omitting participant recruitment information would "fail" this criterion in the CHERRIES electronic survey reporting checklist (Eysenbach, 2004), lowering its perceived quality for reviewers knowing about CHERRIES.

Despite the thousands of specific quality-related criteria used to judge academic outputs, three general dimensions cover the most common aspects of quality across academia: originality/novelty/innovation, plausibility/reliability/rigour/solidity, and likely value/usefulness/significance/impact (Langfeldt et al., 2020). The last dimension can be split into academic and societal value (Aksnes et al., 2019). Although the three core concepts seem to be widely accepted across disciplines, perhaps helped by national research evaluation exercises that mention them, they are interpreted differently (Hamann and Beljean, 2019). These criteria are sometimes explicitly graded, such as nationally vs. internationally relevant (Bonaccorsi, 2018, p. 82), with simple methodological robustness criteria for meta-analyses (e.g., Key et al., 2006) and with hierarchies of evidence in medicine (Blunt, 2015). Grammar and formatting issues may be regarded as aspects of the quality of academic work but are less fundamental.

Previous studies of why academics might make substandard judgements about research quality for journal articles or grant proposals have not tended to compare reviewers. Nevertheless, all the issues found seem likely to vary between reviewers, including cognitive

particularism (preferring one paradigm or topic) (Travis and Collins, 1991), gender/nationality/language/prestige/confirmation bias (Lee et al., 2013), anti-innovation bias, cronyism, the burden of peer review (Guthrie et al., 2017), and even bias *against* useful and clearly written research (Tourish, 2020). For grants, the difficulty in reliably assessing applications has even led to calls for formal randomness to replace the implicit randomness of the decision process (Horbach et al., 2022).

Measuring inter-reviewer consistency

A logical way to assess the robustness of journal peer review is to compare reviewers' recommendations since there is no "gold standard" measure of academic quality. There is also no measure of peer review quality (Garcia-Costa et al., 2022), although there are checklists for the types of information contained in a report (Superchi et al., 2020; Van Rooyenet al., 1999), which is useful for simple tests (e.g., Schroter et al., 2006). Ideally, reviewer recommendations will always be the same because reviewers are expert enough to perfectly evaluate a paper based on a tacitly or explicitly agreed concept of quality. Crucially, any differences between reviewers suggests that issues caught by only one reviewer could easily have been overlooked altogether if a different reviewer had been selected for the team. In practice, as previous studies have shown (Table 1), the norm is substantial disagreement between reviewers, sometimes including fundamental disagreements on appropriate components of quality or acceptable goals and methods (e.g., Sheard, 2022).

There are multiple formulae to assess inter-reviewer consistency, which typically involves checking inter-rater reliability for ordinal data. These formulae include Pearson/Spearman correlation (these are the same for scores ordered as ranks), intra-class correlation (Bartko, 1966) and Cohen's Kappa (weighted, if the outcomes are not binary) (Cohen, 1960). Of these, the first is suitable for paired data (two reviewers per output), the second can accommodate varying numbers of reviewers and the third can be used when the same set of reviewers checks each article. The standard metric for assessing inter-reviewer reliability is intraclass correlation (ICC), although Finn's r (Finn, 1970) is better for some data (Tinsley and Weiss, 1975). ICC is a variant of the Pearson correlation which does not assume that the two sets of reviewer scores have the same mean and variance (Koo and Li, 2016; Shrout and Fleiss, 1979).

Table 1. Intraclass correlation ICC(1,1), Pearson's r, interval, and Cohen's Kappa for overall journal recommendations. Categories range from Accept to Reject (various wordings and minor variations), unless stated. Kappa is weighted unless stated.

Journal	ICC	r	Kappa	Articles	Scale	Reviewers	Source
Health sciences review	0.71			36	7	9 mixed level judges	Oxman et al. (1991)
American Psychologist	0.59			71	4	Journal	Hargens & Herting (1990b)
Stroke	0.55			<12,902	3	Journal	Sposato et al. (2014)
American Psychologist	0.54		0.52	87	5	Journal	Cicchetti (1980)
Developmental Review	0.44			72	4	Journal	Whitehurst (1983)
Research Quarterly for Exercise and Sport	0.37		0.11	363	4	Journal	Morrow et al. (1992)
South African Journal of Psychology	0.34			164	4	Journal	Plug (1993)
American Sociological Review	0.28			322	4	Journal	Hargens & Herting (1990b)
Physiological Zoology	0.28			209	4	Journal	Hargens & Herting (1990b)
Journal of Counseling Psychology	0.28			207	4	Journal	Munley et al. (1988)
Journal of Personality & Social Psychology	0.26			286	3	Journal	Scott (1974)
Atmospheric Chemistry and Physics**	0.24		0.22	356	4	Journal	Bornmann & Daniel (2010)
Personality & Social Psych. Bulletin	0.23			177	5	Journal	Hargens & Herting (1990b)
Clinical Rehabilitation	0.20			193	11	Journal	Wade & Tennant (2004)
Journal of Abnormal Psychology	0.19		0.15	1067	4	Journal	Cicchetti & Eron (1979)
"Major subspecialty medical journal"	0.17+			866	4	Journal	Cicchetti & Conn (1978)
Law and Society Review	0.17			251	4	Journal	Hargens & Herting (1990b)
Lancet		R=0.83		141	6	Journal	Marušić et al. (2002)
Croatian Medical Journal		R=0.78		140	6	Journal	Marušić et al. (2002)
Int. J. of Social Work Values & Ethics		R=0.75		440	5	Journal	Marson & Lillis (2022)
Angewandte Chemie			0.43	718	4	Journal	Bornmann et al. (2008)
Anonymous clinical neuroscience journal			0.28*	116	3	Journal	Rothwell et al. (2000)
Physical Therapy			0.11*	223	3	Journal	Bohannon (1986)
Anonymous clinical neuroscience journal			0.08*	179	3	Journal	Rothwell et al. (2000)
Physical Therapy			0.01*	223	5	Journal	Bohannon (1986)

*Unweighted Kappa (i.e., classes treated as categories) **Open Access journal. +calculated separately for each class, with values between 0.17 and 0.40.

The most appropriate ICC variant for journal peer review is ICC(1,1), for different reviewers assessing a range of outputs. ICC(1,1) is the variance due to differences between papers divided by the total variance due to differences between papers and differences between reviewers (Liljequist et al., 2019). The ICC(1,1) formula is as follows, where k is the number of reviewers (often two in a study), MSBP is the mean square between papers (i.e., the mean of the squares of the differences between the average score for each paper and the overall average score for all papers) and MSBR is mean square between reviewers for the same paper (i.e., the mean of the squares of the differences between the reviewer scores for a paper and the average score for that paper) (Liljequist et al., 2019).

$$ICC(1,1) = \frac{MSBP - MSBR}{MSBP + (k-1)MSBR}$$

Empirical results about reviewer disagreement should be interpreted cautiously because of the following factors, amongst others.

- The extent to which they agree on the merits of the paper evaluated. This is what ICC is designed to measure, but the factors below also influence the calculation. This main reason may be influenced by the training, expertise, or time taken by the reviewers as well as the inherent difficulty of the reviewing task.
- Editors might choose reviewers for diverse perspectives or to address different facets of a submission (Hargens and Herting, 1990a). This would produce lower agreement rates without being evidence of reviewer inconsistency. No studies reporting ICCs have included information on this facet, perhaps because the extent to which it occurs varies between submissions, so it has an unknown influence.
- The degree to which the two reviewers interpret the rating scale (e.g., numerical scores or accept/minor revisions/major revisions/reject) in the same way. This is unknown and not reported in ICC studies. If instructions are unclear or absent, then this would produce lower agreement rates that would exaggerate the extent to which reviewers inconsistently judged the content. This issue partly involves the reviewer's judgement about the standards of the journal as well as their understanding of the scale.
- The extent to which the submissions evaluated have relatively uniform quality. It is harder to get a high ICC(1,1) after a strict desk rejection or other quality filtering (Erosheva et al., 2021), including through scholar self-filtering for quality (e.g., if most scholars in a field choose a journal based on the perceived quality of their work). Conversely, journals attracting submissions with highly varied quality would tend to have higher ICC(1,1) scores for the same level of underlying reviewer consistency. No previous studies have reported desk rejection rates and, in any case, these rates are not directly comparable between journals because there are presumably differences in the percentages of low quality submissions to each one.
- The range of scores that reviewers in the field feel appropriate to give to the papers. For example, if they avoid giving the lowest score for politeness reasons or the highest score on the principle that no research is perfect, then this would reduce the range of scores given and hence ICC values.
- The coarseness of scheme used when decisions are often marginal. ICC calculations are designed to factor out the coarseness of the scheme, so a ten-point scale ICC would be comparable to a three-point scale ICC, in theory. Nevertheless, if decisions are often marginal (e.g., articles are often on the border between minor and major revisions) then a more fine grained scheme would tend to give higher ICC(1,1) scores.

In summary, comparisons between ICC(1,1) scores (or between other inter-rater reliability metrics) only assess the extent to which the selected reviewers agree on their understanding of the scale used and also depend on the quality of the sample of papers examined. ICC values are never fully comparable between studies – or even between journals in the same study - because of the unknown factors that influence them. Nevertheless, it is useful to compare them to identify potential trends across studies, with the hope of making tentative conclusions. In general, more robust studies of inter-rater reliability (e.g., larger sample sizes) have tended to report lower reliability coefficients (Bornmann et al., 2010), however, which complicates comparisons further.

As a corollary to the above, reliability metrics for published articles assess the extent to which reviewers agree on a set of good articles, whereas reliability metrics for a mixed set that includes published and rejected articles at least partly assesses the extent to which reviewers can distinguish between acceptable and unacceptable research.

Inter-reviewer consistency for quality aspects of journal articles

Several studies have reported inter-reviewer consistency scores for quality-related aspects of reviews, usually exploiting scores submitted as part of the reviewing process, although none have assessed the three core quality dimensions for journal articles. One reported low ICCs for pairs of reviewers for 356 articles in the open access journal *Atmospheric Chemistry and Physics*: significance incorporating originality (0.33), methods and results (0.27), and presentation (0.34) (Bornmann and Daniel, 2010). Low ICCs have also been found for 52 *Developmental Review* articles for likely impact (0.23) and originality (0.11) (see also below) (Whitehurst, 1983). Although a conference rather than a journal, even lower ICCs were reported for 145 submissions to an unnamed interdisciplinary conference, presumably rating substantial full texts: relevance (0.21), novelty (0.28), significance (0.17), and soundness (0.21) (Jirschitzka et al., 2017). Thus, it seems that inter-reviewer agreement for quality dimensions is likely to be low, although these results do not prove that there are no contexts in which it can be high. In particular, the two correlations above related to rigour/validity are very low (0.27 for methods and results and 0.21 for soundness).

Inter-reviewer and inter-judge consistency for detailed journal requirements

A few projects have exploited scores submitted by the original reviewers or subsequently chosen judges (although "judges" review articles, different terminology is used to differentiate them from the usual editorially selected "reviewers") for detailed aspects of journal articles to check for consistency between reviewers (one paper also asked the authors: Aksnes et al., 2023). These aspects were presumably chosen by the editor to help guide the overall decision or to direct reviewers to think about aspects of submissions considered important by the editor. In theory, agreement scores for these should be higher than for overall quality criteria because they are more specific, so there is less scope for ambiguity about what should be assessed. These are discussed below separately for judges assessing multiple articles and for standard journal reviewers. Unlike the current article, none have assessed agreement separately on the three core dimensions of research quality.

Table 2. Intraclass correlations ICC(1,1) between quality aspect scores from reviewers of several journals.

Journal	Validity	Significance	Originality	Clarity	Articles	Source
Social Work Research	most important studies cited (0.345), suitable research design (0.404), detailed methods description (0.128), detailed statistics description (0.127), correct statistics (-0.015), data support conclusions (0.175)	important for social work and welfare (0.175)	new information or justified replication (0.283)	clear, succinct, and organised article (0.279)	34-54	Kirk & Franke (1997)
Canadian Journal of Behavioural Science	literature review (0.44), design (0.27), analysis (0.31), interpretation (0.34)	importance (0.16), appropriateness for the journal (0.21),		presentation (0.23)	120	Linden et al. (1992).
Journal of Counseling Psychology	relation to literature (0.35), methodology (0.28), interpretation of results and conclusions (0.19)	significance of topic (0.20), importance (0.28),		presentation clarity (0.24), length (0.13)	232-262	Munley et al. (1988)
South African Journal of Psychology	theory related to analysis (0.42), literature review (0.35), research design and application (0.33), methods (0.23), results interpretation (0.12)	contribution to psychology (0.28),	theory developed (0.28)	structure (0.11), language (0.15), author guidelines followed (0.25)	64-117	Plug (1993)
Journal of Personality and Social Psychology	literature review (0.37), design and analysis (0.19)	problem interest (0.07), importance (0.28),		style and organisation (0.25), and succinctness (0.31)	312-574	Scott (1974)
Developmental Review	conceptualisation (0.17), theory (0.43), validity of conclusions (0.21)	likely impact (0.23), topic importance to	originality (0.11)	writing (0.28)	52	Whitehurst (1983)

		field (-0.10), wide interest to developmentalists (0.17)				
Journal of Abnormal Psychology	research design (0.33), literature review (0.29), data analysis (0.24)	importance (0.23), reader interest (0.20)		style/ organisation (0.18), succinctness (0.29)	356-400	(Cicchetti & Eron, (1979).
Clinical Rehabilitation	method measures (0.25), method design (0.21), method analysis (0.26), results tables/figures (0.33), discussion weaknesses (0.19), and discussion extrapolation (0.13)			readability (0.20), abstract (0.22), introduction (0.12), method description (0.27), results presentation (0.22)	193	Wade & Tennant (2004)

Inter-*reviewer* consistency for journal requirements

Reviewer agreement on specific aspects of journal articles tends to be low (Table 2). This may be due to difficulties obtaining satisfactory reviewers for journals (e.g., perhaps including inexperienced reviewers), the lack of care with which reviewers rate articles (e.g., perhaps assuming that the text of the review and the overall judgement are the only important things), or the ambiguity of the task of assigning a grade to a submission as a one-off exercise without other articles to grade at the same time to give context. Thus, the low agreement rates may reflect task uncertainty rather than genuine disagreement. In support of this, grant reviewers tend to be more consistent when they have previous experience of the same funding scheme, so having prior knowledge to norm reference against but reviewing experience in general does not improve consistency (Seeber et al., 2021). Nevertheless, it is not clear whether agreement rates for journal article reviewers would be higher in the ideal situation that all reviewers could fully understand all aspects of the article evaluated.

Inter-*judge* consistency for journal requirements

At least four studies have used judges other than journal reviewers to assess journal requirements for a set of articles. Nine junior, average or expert reviewers assessed each of 36 health sciences review articles on nine methods quality dimensions and overall, achieving high ICC scores: search methods reporting: 0.87, search comprehensiveness, inclusion criteria reporting: 0.86, avoiding selection bias: 0.68, reporting validity criteria: 0.68, appropriate validity assessment: 0.72, combining methods reported: 0.62, appropriately combined findings: 0.52, data supports conclusions: 0.40, overall scientific quality: 0.71 (Oxman et al., 1991). This was an unusual task, however, assessing only the methods validity of review articles, a genre with broadly accepted criteria in the health sciences, and with the overall decision presumably influenced by the chosen methods criteria rated.

Fifteen judges of different types have rated 36 articles describing pain-related clinical trials on 11 different methods aspects (e.g., “Were the outcome measures clearly defined?”) (Jadad et al., 1996) but these are too specific to relate to quality. Totalling the answers gave high intraclass correlations ICC (0.60 for all 15 judges, 0.69 for just the four researchers), presumably due to the specificity of the questions. In another checklist-based study, 7 raters (authors of the article) of 227 surgical endoscopy education article methods reached almost perfect ICC agreement (0.96) (Anderson et al., 2022). These high agreement rates suggest that specific defined criteria are more likely for narrow defined quality criteria.

Eight judges (the reporting paper’s authors) assessed the same four empirical software engineering papers in nine dimensions, achieving variable but mostly high levels of agreement (ICC between 0.18 and 0.84, with 0.89 for the sum) (Kitchenham et al., 2012). These high scores may be due to the experimental set up (article co-authors being the judges, only four papers assessed).

Overall, this small set of studies suggests that a high degree of consistency can be achieved when a common set of judges assess the quality of journal articles with narrowly defined criteria in the health domain, and perhaps others.

Methods

Data

All data was obtained from the SciPost website that hosts the SciPost Physics online journal. The SciPost sitemap https://scipost.org/sitemap.xml was downloaded on 23 August 2022. This apparently lists all pages in the site, including unpublished articles. From this list, all URLs ending in "v[number]/", where [number] denotes the manuscript version number, were selected (e.g., https://scipost.org/submissions/2202.11102v3/). These were downloaded at a rate of 1 per 10 seconds 23-24 August 2022, in ethical compliance with the site robots.txt file.

Submission URLs end in version numbers, usually starting at 1 but with the URLs otherwise being the same for all versions. Based on these version numbers, the earliest posted review was selected for each article and the remainder were discarded, leaving 2375 article review pages. The first available review was usually the only one with review scores, with the later reviews on revised versions of the original submission usually containing just text. The sample is thus first round reviews. The dataset includes reviews for papers that have been accepted and those that have been rejected.

A program was written to extract the text of the reviews and the review scores from each first version, and this was added to Webometric Analyst (Services menu, Peer Review submenu, SciPost menu item), which is free online (http://lexiurl.wlv.ac.uk/). The first two reviews for each article were selected for RQ1 and RQ4, and all reviews were selected for RQ3 (focusing on individual reviewers, so the fact that an article is difficult to review and might require extra reviewers is less important). Less than a quarter of articles had three or more reviews, so whilst adding extra reviewers would slightly increase the amount of data available for RQ1, the calculations would be less independent since all reviewers might concur in simple cases. For RQ1 the first two reviewers were selected rather than choosing two reviewers at random because a third reviewer was rare, suggesting that they might be called upon by the editor if the first two reviewers disagreed and so would be non-standard choices in this respect. The average scores for all reviews for the journal were almost the same for the first three reviewers, suggesting that there was little difference between them in practice, however (4.9 for originality for all three; 4.4 for Significance for all three; 4.4 for originality for reviewers 1 and 2, but 4.3 for reviewer 3).

Each review has four text components, but only the third is compulsory: Strengths, Weaknesses, Report, and Requested Changes. Reviewers can optionally also score each article in four quality dimensions: Validity, significance, originality, clarity. These use a six-point scale that is not defined or explained in the site but is presented in order as a drop-down list: poor, low, ok, good, high, top. Reviewers are also requested to score articles for formatting and grammar on a seven-point scale: mediocre, below threshold, acceptable, reasonable, good, excellent, perfect. Although the two scales are non-standard and the order of the labels might not be self-evident from their names (e.g., poor vs. low), their ordering is clear from their positioning in the drop-down lists. All this information is public, but the overall recommendation (a seven-point scale) and the reviewer's self-reported qualifications are private. All reports analysed were labelled "invited report", suggesting that the reviewers had been invited by the editor, rather than volunteering on the site. The reviews can be posted online ahead of a decision within a round, so earlier reviews could affect subsequent reviews if read by the reviewers. Data from another publication platform suggests that prior publication of reports does not affect subsequent reviewers' recommendations, however (Thelwall et al., 2021). This lack of influence could be because reviewers rarely read others'

comments (which we think likely), are rarely influenced by them, or positive influences counteract negative influences (wanting to disagree with a previous reviewer). The potential availability of a prior report for one reviewer is still an important limitation.

The SciPost website hosts multiple journals, but SciPost Physics (n=1615) is the largest. Authors may specify the approach used in their article and the most common is Theoretical. Fortunately, this most common category also seems most likely to be fully understandable to a reviewer. An experimental article, in contrast, may have experimental components opaque to some physicists but statistical components opaque to others. This might lead to different scores due to focusing on different aspects of an article. The likelihood of this seems to be lower for theoretical physics, although not absent since different types of theory might be merged into one article or a theoretical physicist might be asked to review an article incorporating theory that they were not familiar with.

Italian research evaluation reviewer score data for comparison was obtained from Zenodo (zenodo.org/record/4848684) for a sample 7,667 Valutazione della Qualità della Ricerca (VQR) publications from 2011-2014 that were independently evaluated by selected field specialist reviewers. Expert reviewers selected by the VQR gave journal articles scores on a scale of 1-10 for each of significance, originality, and rigour (Traag et al., 2020). Articles in this set typically have scores from two independent researchers possibly evaluating about five outputs each (based on a previous exercise: Minelli et al., 2008). Unfortunately, separate significance, originality, and rigour scores are not available but the sum of the three is: a value in the range 3-30. The total score for each VQR reviewer is therefore conceptually similar to the sum of the SciPost Physics Originality, Significance and Validity SciPost scores.

Analysis

For the first two research questions, intraclass correlations (ICC) were calculated, and more specifically ICC(1,1) since the reviewers are largely different for each article and each give a separate score. As argued above, this is preferable to existing alternatives and was also used in the most similar prior work (Bornmann and Daniel, 2010; Jirschitzka et al., 2017; Whitehurst, 1983), so is easiest to compare with their results. ICC(1,1) was calculated with the ICC() command in the R package psych, version 2.2.5.

Spearman correlations were used for RQ3. Because the scores are also ranks, Pearson and Spearman correlations are the same. The correlations assess the extent to which one quality aspect tends to score higher when the one it is compared to also scores higher. Unlike ICC, it does not assess the extent to which the scores are the same.

To address RQ4, the reviews for disagreeing reviewers 1 and 2 were read and analysed to infer why they may have scored an article at least two points differently, then the results were clustered into themes. This is an ad-hoc type of reflexive thematic analysis (Braun and Clarke, 2019). A formal content analysis (Neuendorf, 2017) would be inappropriate because the sample sizes are relatively small and the journal relatively specialist and unusual. The value of the results is therefore exploratory: revealing the types of reason that occur rather than estimating their frequency for theoretical articles in the SciPost Physics journal.

The main criticism of the lower scoring reviewer was first identified by reading their text related to originality, validity, or significance, and recording the main points, if any. Both authors of the current article completed this stage independently, recording the answer as a short phrase. The first author is an experienced reviewer with a PhD in pure mathematics and a co-author of several physics articles but not a physicist. The second author is a senior theoretical physics professor and experienced interpreter of physics reviews as editor of the

prestigious *Physica A* journal, although not familiar with all physics theories. Thus, both authors have relevant expertise for analysing theoretical physics reviews and the second author is particularly experienced at it. After this stage, the first author checked the reasons identified by both authors and attempted to resolve any differences and fit the reasons into a more general themes, repeatedly checking similar themes against each other for opportunities to merge similar ones together or split large themes into coherent subthemes.

To illustrate the final themes, within the validity set several reviewers seemed to give lower scores after challenging one of the assumptions of the article, so "Invalid assumptions" was selected as a theme. Similarly, within the Originality set, many reviewers stated that the findings had been shown before, and this reason was subsumed within the theme, "Duplicates prior work".

Results

RQ1: Inter-reviewer consistency

There were 505 SciPost Physics (Theoretical) articles with at least two scoring reviewers, with up to 1007 individual reviewer scores for one facet from within this set (Table 3). Just under half of the reviewers gave the second highest score for each facet (5 or 6, depending on the facet), and Validity (28%) had a relatively high proportion of reviewers giving it the top score compared to Significance (9%) and Originality (10%) (Figure 1).

Table 3. Descriptive statistics for individual reviewer scores (top 3 data rows) and pairs of reviews for the same article (bottom 3 rows).

	Validity	**Significance**	**Originality**	**Clarity**	**Formatting**	**Grammar**
Reviewer scores	1001	1007	1005	1005	997	995
Mean score	4.94	4.42	4.41	4.59	5.77	5.82
Score std. dev.	1.06	1.00	1.02	1.16	1.16	1.23
Pairs of scores	497	502	500	500	492	490
Mean difference	0.70	0.73	0.77	0.84	0.77	0.69
Difference std. dev.	0.65	0.65	0.67	0.76	0.70	0.67

The first two reviewers agreed between 40% and 48% of the time for each facet and, when they disagreed, it was usually by one point (39% to 46%) (Figure 2).

The six quality score ICCs for the first two reviewers are all between 0.32 and 0.44 (Figure 3). Despite the large sample sizes, except for the low correlation for formatting, the other quality aspects have overlapping 95% confidence intervals, so it is not reasonable to draw conclusions about which is strongest. The moderate correlations are perhaps surprising given that reviewers seem likely to be usually able to evaluate all aspects of a theoretical physics article since it is based on mathematical calculations. In terms of the six quality aspects, disagreements may be partly due to different interpretations of the overall scale. In particular, some reviewers may have a stricter temperament than others, and reviewers may norm reference against different levels: their own work, the journal submitted, or well-known physics articles. Reviewers may also put different amounts of time and care into a review. The following additional facet-specific issues may also be relevant.

- **Grammar**: Non-English speaking reviewers might be unable to fully evaluate this and English speakers may have unrealistic or different expectations.

- **Formatting**: This is a vague descriptor and might include article structure, image formats, or overall format of the document reviewed so the disagreement may reflect a lack of certainty about what the score represents.
- **Clarity**: Variations in the understandability of the exposition in the article may be due to reviewers with different specialisms and levels of experience.
- **Originality**: This depends on how well the reviewer knows the subject area and an inexperienced reviewer may not be able to judge this. Originality could refer to different aspects: methods, application area or exposition.
- **Significance**: As for originality, experience and knowledge may be needed to know whether a given article is likely to be significant to the community. Significance may also be interpreted as meaning internal to physics, internal to academia, or societal relevance.
- **Validity**: This seems the most straightforward aspect of an article to judge but subjectivity arises because, other than through mathematical proofs, arguments rest on an accumulation of evidence. Thus, a validity judgement may entail assessing the strength of the evidence provided.

Figure 1. Distribution of quality facet scores from the first two reviewers for the 505 SciPost Physics articles classified by their authors as Theoretical, and having at least two sets of reviewer scores, each with at least one non-missing score. There is no score 7 for the first four facets: these are on a six-point scale.

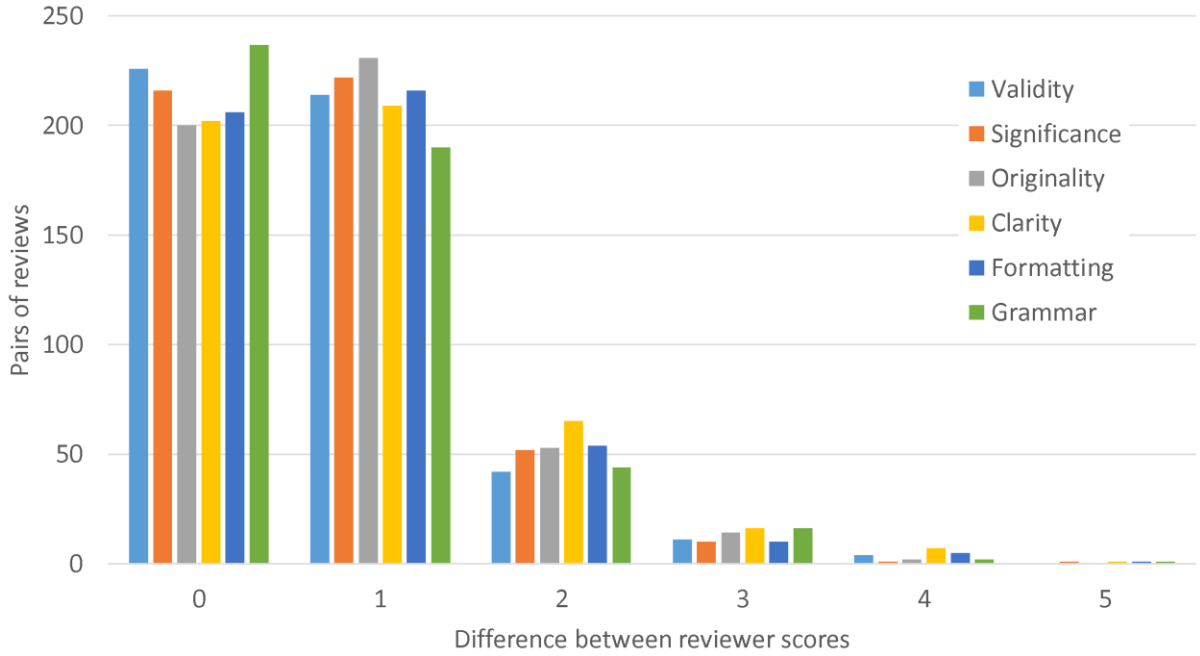

Figure 2. Distribution of quality facet score differences between the first two reviewers for the 505 SciPost Physics articles classified by their authors as Theoretical, and having at least two sets of reviewer scores, each with at least one non-missing score.

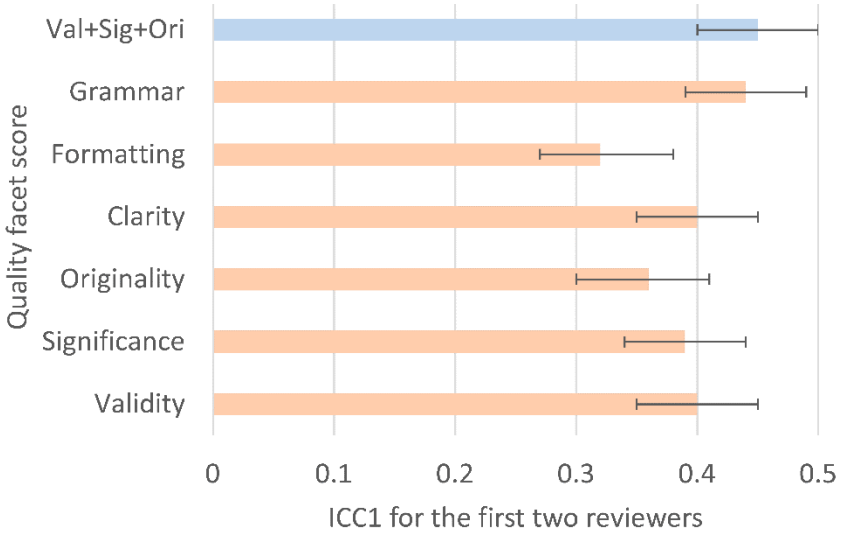

Figure 3. Intraclass correlations ICC(1,1) between quality facet scores from the first two reviewers for the 492- 502 SciPost Physics articles classified by their authors as Theoretical, and with two sets of reviewer scores. The ICC for the sum of the main three quality components is also shown. Error bars show 95% confidence intervals.

RQ2: Inter-reviewer consistency in comparison to the VQR

There were 1008 VQR Physics articles with two scores, with half (51%) of the validity + originality + significance totals in the range 21-27 (Figure 4). The distribution shape is on a finer scale but otherwise similar to that of the sum of validity, originality and significance for the 505 SciPost Physics theoretical articles (Figure 5), in the sense that medium and low scores are rare, as are the highest scores in both cases. Nevertheless, the VQR distribution is bimodal

whereas the SciPost Physics distribution is unimodal, perhaps suggestive of reviewers deliberately avoiding giving maximum scores in the latter case.

VQR reviewers' total scores agreed 9% of the time, were within 3 of each other 51% of the time, and were within 9 of each other 89% of the time (Figure 6). This distribution is similar in shape to the SciPost reviewer difference distribution for validity + originality + significance totals (Figure 7). The two VQR Physics distributions (Figure 4, Figure 6) are close to the equivalent distributions for the other VQR fields (not shown).

The ICC for the totalled quality scores of the first two reviewers for SciPost theoretical physics articles (Figure 3) is slightly higher than the ICCs for totalled quality scores for all VQR subject categories (Figure 8), although the differences are within a margin of error in most cases. Most importantly, comparing the SciPost theoretical physics overall ICC (0.45) with the VQR physics category ICC (0.40), the difference is small (0.05). The two tasks (journal reviewing, post-publication output scoring for national research evaluation) are not directly comparable, however. In theory, quality should be more variable for SciPost since it includes some rejected articles, which should increase its ICCs by increasing the number of low scores that reviewers agreed on. The score distributions for the VQR and SciPost totals (Figure 4, Figure 5) show the opposite however: there is a smaller percentage of low scores for SciPost than for VQR. Thus, for whatever reason, the VQR ICC has a technical advantage over the SciPost ICC, which makes the slightly higher SciPost ICC more impressive by comparison. Aside from this technical reason, factors that might lead one ICC to be higher than the other include the following.

- **SciPost greater inter-reviewer reliability than VQR**: Reviewers must write a public report justifying their scores. Reviewers of theoretical physics outputs may be more likely to evaluate all aspects of them, avoiding conflicts due to primarily evaluating different parts (e.g., theory, methods, statistics).
- **VQR greater inter-reviewer reliability than SciPost**: Reviewers may be more consistently senior and expert, since their selection has to be validated by a panel. The VQR drives funding and is more important for careers. VQR reviewers are presumably given instructions to help interpret the scales. Since VQR reviewers carry out multiple reviews (e.g., five: Minelli et al., 2008), they can norm reference their scores to some degree.

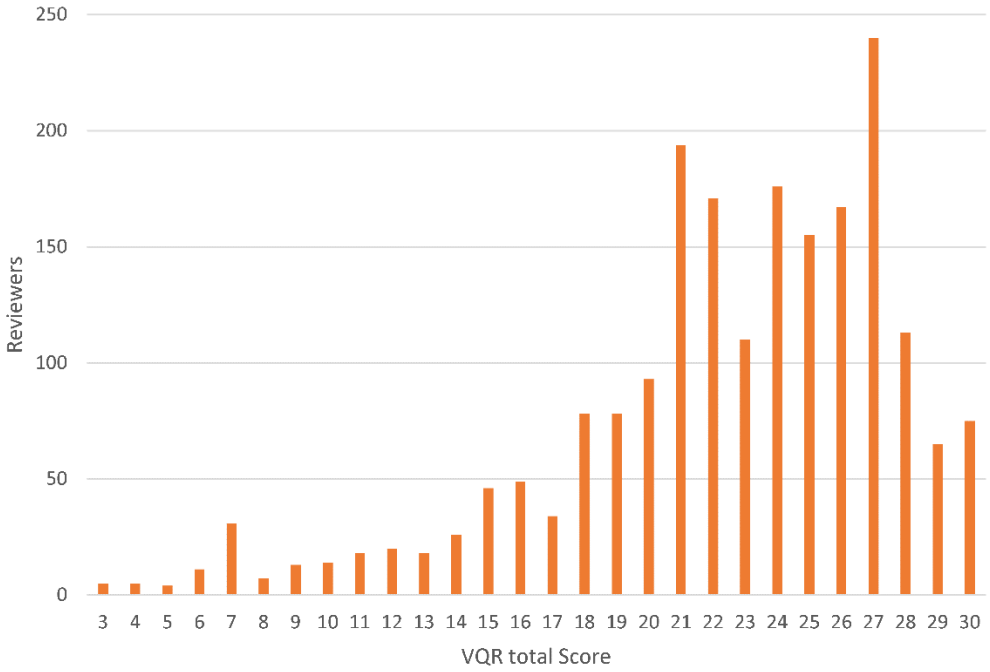

Figure 4. Distribution of quality facet score totals (i.e., validity +significance+originality) from the two reviewers for the 1008 VQR Physics articles.

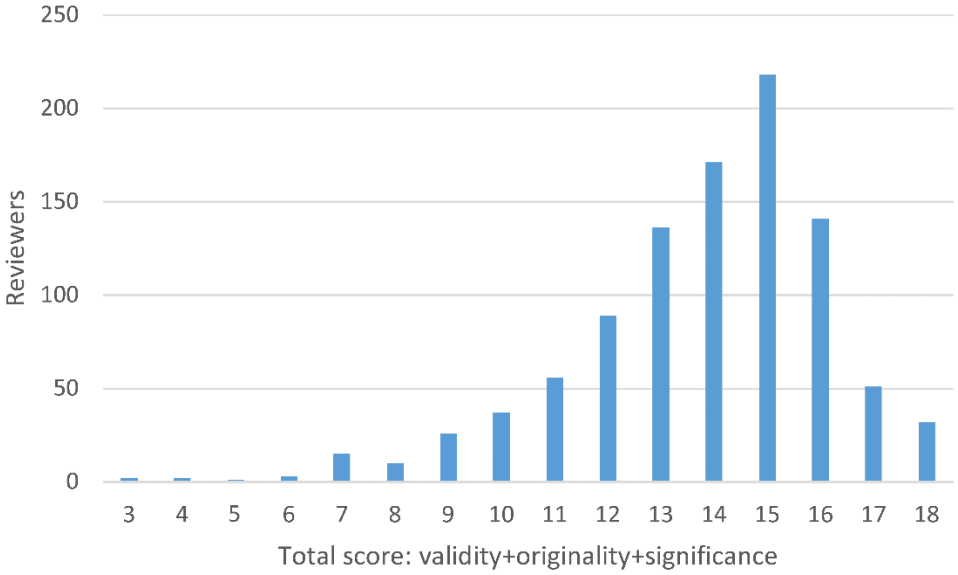

Figure 5. Distribution of quality facet score totals (i.e., validity +significance+originality) from the two reviewers for the 505 SciPost Physics theoretical articles.

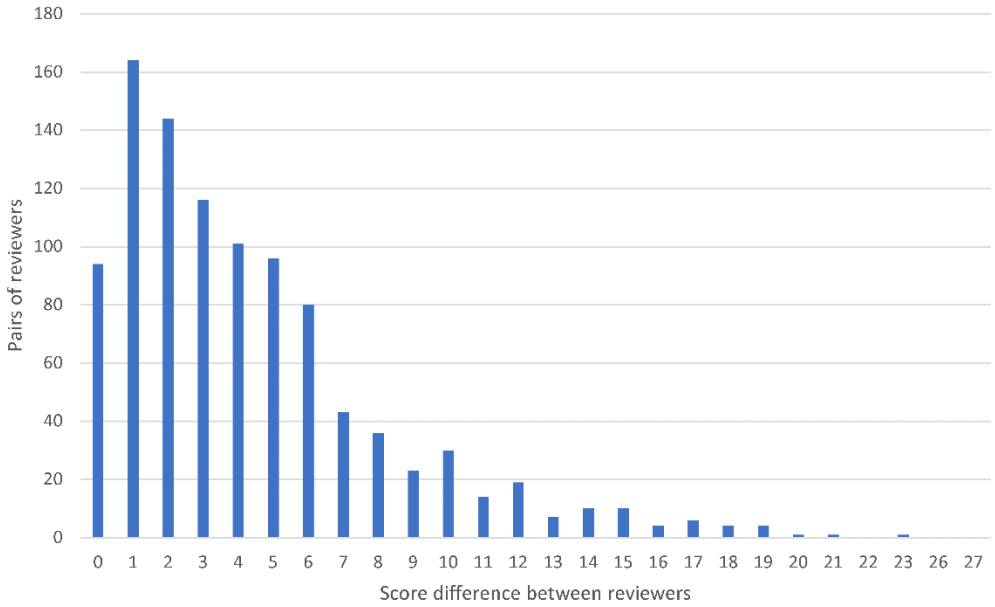

Figure 6. Distribution of quality facet score differences between the two reviewers for the 1008 VQR Physics articles.

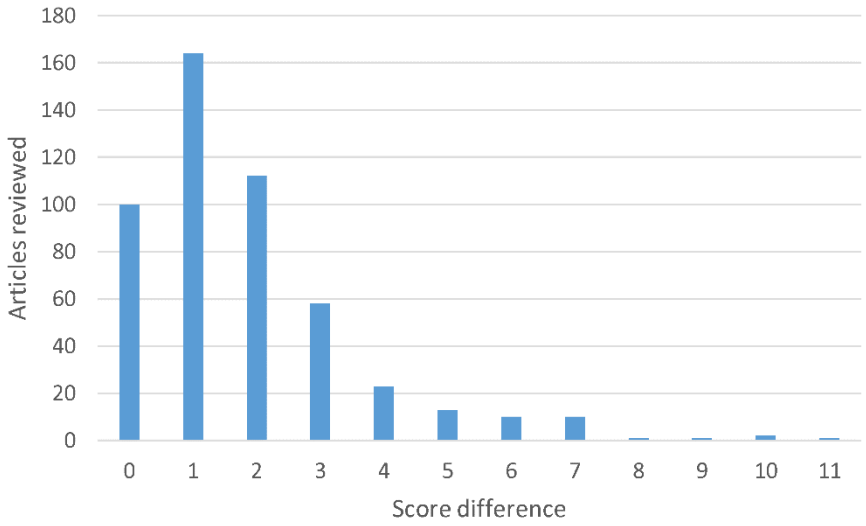

Figure 7. Distribution of quality facet score differences between the two reviewers for the 505 SciPost Physics theoretical articles.

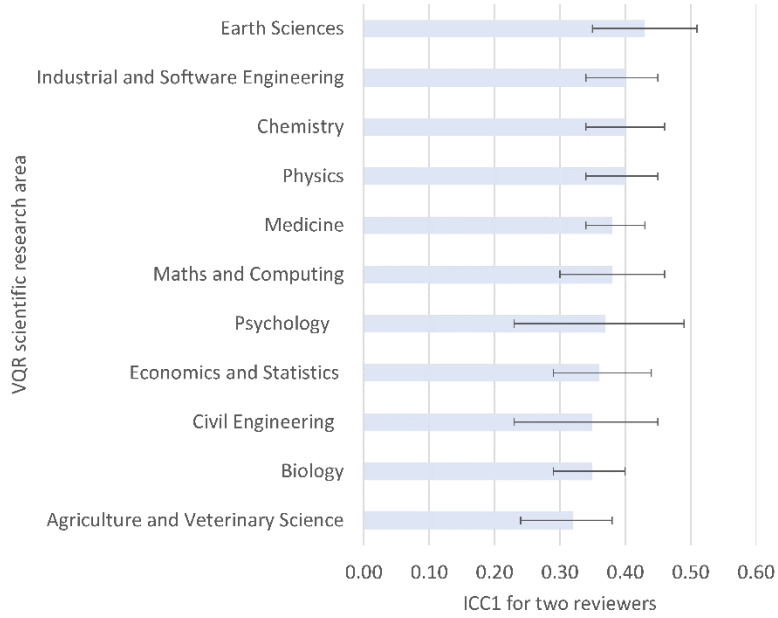

Figure 8. Intraclass correlations ICC(1,1) for quality facet scores totalled (i.e., validity +significance+originality) from two reviewers for selected VQR outputs by subject category. Error bars show 95% confidence intervals. Sample sizes are between 175 and 1293 articles.

RQ3: Facet score similarity for the same reviewer

There are substantial differences in the extent to which reviewers assign similar scores to different quality aspects (Figure 9). For the three main quality criteria, Significance and Originality appear to be the most related aspects and Validity-Originality the least. In theory, both rigour and originality are important for significance, but the correlation between rigour and significance may be reduced by robust but routine studies that contribute little to the scholarly record. These might be demonstrations that known methods work in a slightly different context to previously shown, for example. It is not possible to assess whether the judgements are fair in the sense that reviewers sometimes penalise one aspect (e.g., Validity) for low scores on another (e.g., Grammar), or give all aspects of an article the same score without much consideration.

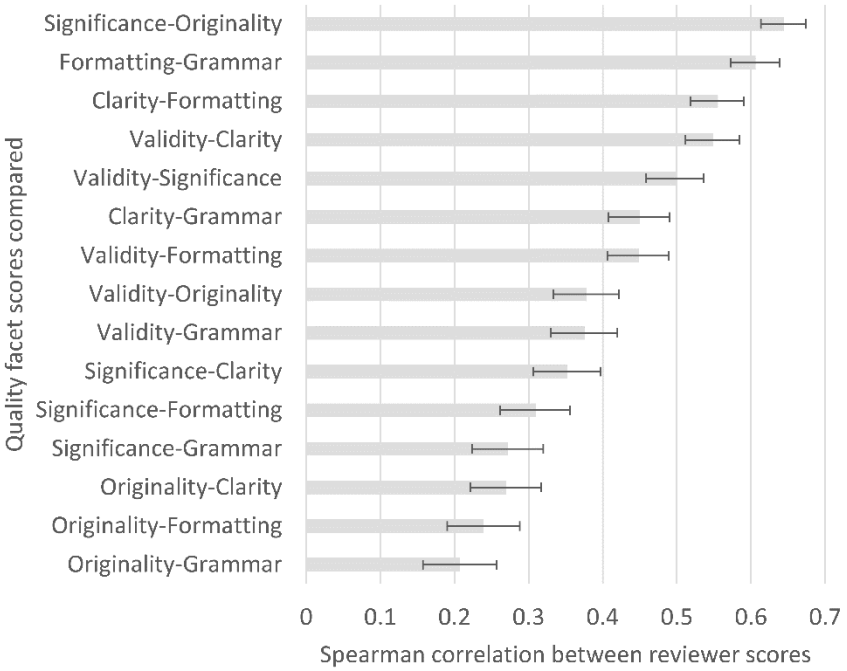

Figure 9. Spearman correlation between quality facet scores from all reviewers for SciPost Physics articles classified by the authors as Theoretical. Error bars show 95% confidence intervals. Sample sizes (excluding missing values) are between 1447 and 1443 articles.

RQ4: Reasons for lower quality scores from one reviewer

Five common validity-related criticisms were identified by the analysis of review texts (Table 4). In all cases, the difference between reviewers might be one of differing strictness, norm referencing, or level of care/time/detail given to the review. The reviewers may also take a different perspective: has the author shown enough skill to be rewarded with a publication (i.e., the managerial perspective on research), or will the publication make a valid and useful contribution to the scholarly record (i.e., the knowledge perspective on research)? No validity reasons could be found in 8 of the reviewers' reports, leaving 49 reviews with at least one reason from the list below. In some cases, the higher scoring reviewer mentioned the lower scoring reviewer's main point but may have given it a lower weight (e.g., "The real feasibility of the experiment remains therefore questionable." 3/6 vs. "A more complete analysis of the scheme robustness in realistic setups is missing." 5/6). The second reviewer may have considered this an issue for significance (no applications) rather than validity (does not fit applications).

Table 4. The most common validity-related criticisms identified from the reviewer giving the lower validity score (n=57 scoring at least two grades lower).

Validity category	Harsher criticisms of article	Possible causes of reviewer disagreement
Inadequate explanations (n=22; 45%, e.g., "Some points are unclear")	Confusing/inadequate/missing explanations of assumptions, methods or results, missing definitions; figures difficult to understand.	One reviewer may understand the context better or have a higher tolerance for unclear explanations.
Invalid assumptions (n=19; 39%, e.g., "approximations are very crude")	Assumptions conflict with the literature/are too severe/are not explained/justified/clear or motivated/are incorrect/conflict with practical applications. Entire approach or concept is (thought to be) invalid.	One reviewer may understand the context or literature better.
Insufficient evidence (n=8; 12%, e.g., "lack of a solid mathematical result")	More evidence or checks are needed, such as through comparisons against other published approaches.	One reviewer may expect more substantial evidence to support an argument.
Inadequate literature connection (n=6; 10%, e.g., "improve the comparison of the results [] with the existing ones")	Inadequate or missing comparisons to prior work.	One reviewer may be more widely read in an aspect of the background.
Incorrect results (n=3; 6%, e.g., "(1) is not actually a bound")	Invalid conclusions/methods due to errors, including misunderstanding prior work or misinterpreting the results.	One reviewer may understand the context or literature better or spend more time checking methods (some reviewers stated that they had not checked the working).

Five significance-related criticisms were identified, most of which are indirect (Table 5). Significance clearly overlaps with the other categories because an invalid article is insignificant, and an unoriginal paper is unlikely to have much significance. The major reasons for reviewer disagreement for significance is their judgement about how substantial the contribution is and whether the article describes its significance clearly.

Table 5. The most common significance-related criticisms from the reviewer with the lower significance score (n=63 scoring at least two grades lower, excluding 5 with no reason, so 58 overall).

Significance category	**Harsher criticisms of article**	**Possible causes of reviewer disagreement**
Minor contribution (n=27; 47%, e.g., "Not clear why numerical approach is needed")	Minor originality (so little scope for significance)/not clear what can be learned/limited scope. No advantage over other published methods.	One reviewer may have read more prior work.
Inadequate significance claims (n=21; 36%, e.g., "[no] discussion of the implication or significance of the results")	Inadequate or missing discussion of results/implications; misses the chance to generalise to other contexts; importance unclear; does not relate to other work to show contributions. Poorly motivated.	One reviewer may expect the author to spell out non-academic applications.
Little or no practical application (n=5; 9%, e.g., "It may not apply to the real Josephson junctions")	May not apply to real world problems/irrelevant context analysed. Only very specialist applications/narrow contexts.	One reviewer may consider non-academic applications important.
Too complex for end users (n=3; 5%, e.g., "Too technical")	Too technical.	One reviewer may consider the practicalities of applying a method.
Invalid results (n=3; 5%, e.g., "I [disagree] with the main physics ideas")	Results invalid so not significant.	One reviewer may have checked validity more carefully or with more relevant knowledge.

The three originality-related criticisms found were differences in judgements about how substantial the results were, or knowledge/judgement that the results duplicated prior work (Table 6).

Table 6. The most common originality-related criticisms from the reviewer giving the lower originality score (n=65 scoring at least two grades lower, but 26 did not give a reason, leaving 39).

Originality category	**Harsher criticisms of article**	**Possible causes of reviewer disagreement**
Limited novelty (n=18; 46%, e.g., "The result is nice but not that impressive")	Minor/unsurprising advance/trivial generalisation/limited novelty/straightforward application or standard approach.	One reviewer may expect a more substantial contribution.
Duplicates prior work (n=16; 41%, e.g., "The content lacks novelty")	Result completely/partly published before or equivalent to something published before. Not novel. Mainly combines previous works. Longer version of previous paper. Issue has been discussed extensively before.	One reviewer may have read more prior work.
Novelty claim unclear (n=5; 13%, e.g., "Unclear to what extent the paper contributes to our knowledge on []")	Difficult to work out what is claimed to be novel in the paper/need to compare with cited literature/invalid claim.	One reviewer may have read more prior work and may be better able to detect novelty.

In summary, the reviewers may have different types or levels of relevant background **knowledge** (literature) and **understanding** of the problem. They may also have different and perhaps conflicting **beliefs** about valid approaches. They may have dissimilar **expectations** from a research paper in their field or the specific journal in terms of all the components of a paper. These are in addition to diverse **norm referencing**, and **care** with reviewing.

Discussion

The results have many limitations in validity and significance. They are restricted to a single journal, which may be atypical. Reviewing varies substantially between fields in terms of review length and outcomes (Thelwall, 2022), so results may not generalise well. The characteristics of the reviewers are unknown, as is their selection mechanism, and the extent to which they interpret the score ranges in the same way. Some of the second reviewers may have seen the first reviewer's report before posting their scores. The results from RQ4 are subjective, especially for the named themes identified. It is also not known whether the RQ4 results are specific to differences between reviewers or essentially describe the key aspects considered by any reviewer.

The next few paragraphs discuss the ICC scores found, comparing them with prior research. Recall that ICC scores are not only influenced by the level of reviewer agreement but also by the variety of quality levels in the articles assessed and the other factors mentioned in the bullet point list in the Background section. They are not directly comparable between studies because they depend partly on desk rejection strictness and the uniformity

in quality of the submissions (if a journal) or the uniformity in quality of the researchers (for post-publication expert review, like the VQR). Thus, ICC scores from different contexts are not strictly comparable and all the numerical comparisons below should be regarded as approximate rather than definitive. ICC scores have been routinely compared in prior studies despite this limitation, presumably with the implicit understanding that in social science research full comparability between studies is impossible but that comparisons are still an important part of attempting to generalise results.

For RQ1, the ICCs for originality (0.36), significance (0.39) and validity (0.40) for theoretical articles in SciPost Physics are all higher than the corresponding scores previously found for *Atmospheric Chemistry and Physics* (Bornmann and Daniel, 2010), moderately higher than the corresponding scores for an interdisciplinary conference (Jirschitzka et al., 2017) and much higher than the corresponding scores for *Developmental Review* (Whitehurst, 1983), including for the key validity component in all cases. This is consistent with the rationale for this article that theoretical physics reviewers would agree more than average because they could understand all aspects of an article and work within a relatively mature, and hence relatively uncontroversial, field. Nevertheless, disagreement is still the norm for reviewing in this journal, albeit usually by a single score point. As suggested above, the level of disagreement may be partly due to ambiguity over the scoring system for those without prior experience to norm reference their judgements against.

For RQ2, the overall reviewer consistency (summing the three main quality components) for SciPost Physics theoretical articles (0.45) is not only marginally higher than the VQR Physics correlation but is also higher than all except three ICC scores for overall journal article evaluations (Table 2). One of the two exceptions, Stroke, might have high reliability as a medical journal: arguably a mature field with a consensus on reviewing and careful reviewers because of the health implications of the research. It is not clear why American Psychologist reviews would be substantially more consistent, though (Cicchetti 1980; Hargens and Herting, 1990b). It is possibly a statistical anomaly due to the small sample sizes or perhaps the editors select reviewers particularly carefully for expertise, rejected inadequate reviews, or had a lax desk rejection strategy that allowed weak submissions to be reviewed and easily rejected. The four categories used in this case were: accept, accept with minor revisions, revise and resubmit, and reject. The original paper mentioned the journal's requirement for reviews related to controversial issues (Hargens and Herting, 1990b) and this, combined with a focus on a single article type (reviews) may have helped reviewers to make relatively consistent decisions.

For RQ3, no previous study has assessed the extent to which reviewer scores for different quality dimensions for journal articles correlate. One study has checked this for reviewing of papers for an interdisciplinary conference, however, giving: novelty-significance: 0.56, novelty-soundness: 0.41, and significance-soundness: 0.40 (Jirschitzka et al., 2017). These values are lower than for the current paper but in the same rank order, providing support for the ordering here not being peculiar to the journal.

For RQ4, the results do not map easily to previous results about the contents of peer review reports (e.g., Falk Delgado et al., 2019). Moreover, none seem to have analysed physics in the past, and none have qualitatively compared reports between reviewers, although some have qualitatively analysed multiple reports for the same manuscripts (e.g., 32 reports for 4 papers: Sheard, 2022). The comparison between reviews suggests that differences in reviewer scores are not just related to norm referencing but also due to differing identification of problems. It is unsurprising that decisions can be affected by the level and

extent of the reviewer's knowledge and understanding of the article, with different reviewers presumably typically having at least partially non-overlapping areas of expertise. Thus, one reviewer may identify problems with validity or originality that the other did not know about. A reviewer's guess about the significance of a paper is also affected by their understanding of the field and possible related applications. Moreover, some reviewers oppose paradigms or approaches that are nevertheless accepted by others (Whitley, 2000). Nevertheless, based on the literature review (Oxman et al., 1991 vs. the others in Table 1), the substantially lower agreement rates for journal reviewers than for judges rating a large set of articles suggests that norm referencing is the biggest single cause of discrepancies between reviewers in their scores or outcomes. This is credible because validity, significance, and originality are not absolute concepts. Although an article could be almost 100% valid in theory (e.g., a pure mathematics proof), in most areas of scholarship a connection to external reality is needed, and this is not possible to fully test. Related to this, reviewers may have differing levels of expectations about what an author or paper should achieve to be publishable in the journal. This connects with many reviewer comments asking for more extensive testing. Given a lack of guidance from a journal or scores, the reviewer must decide what to norm reference against and might choose Nobel Prize research as the maximum or a good article in the journal. The level of disagreement found is still concerning from a quality control perspective: unless editors usually solicit teams with non-overlapping expertise that collectively assess articles comprehensively, the results suggest that there is a substantial degree of randomness in the peer review system, so that an unlucky choice of referees could allow a good article to be rejected or a bad one published.

Figure 10 summarises how a range of key factors may affect a reviewer's judgement about a journal article, ignoring technical reasons for ICC differences. This depends on many dimensions of their expertise, as well as their personal beliefs (Strevens, 2021), and the care given to reading a paper. After this, various aspects of the paper will be understood to various degrees and the reviewer may form an opinion. This opinion may be translated into a judgment on the journal's scale (e.g., accept, major revisions, reject) through implicit norm referencing either with the reviewer's previous experience or their beliefs about the role of reviewing for the journal.

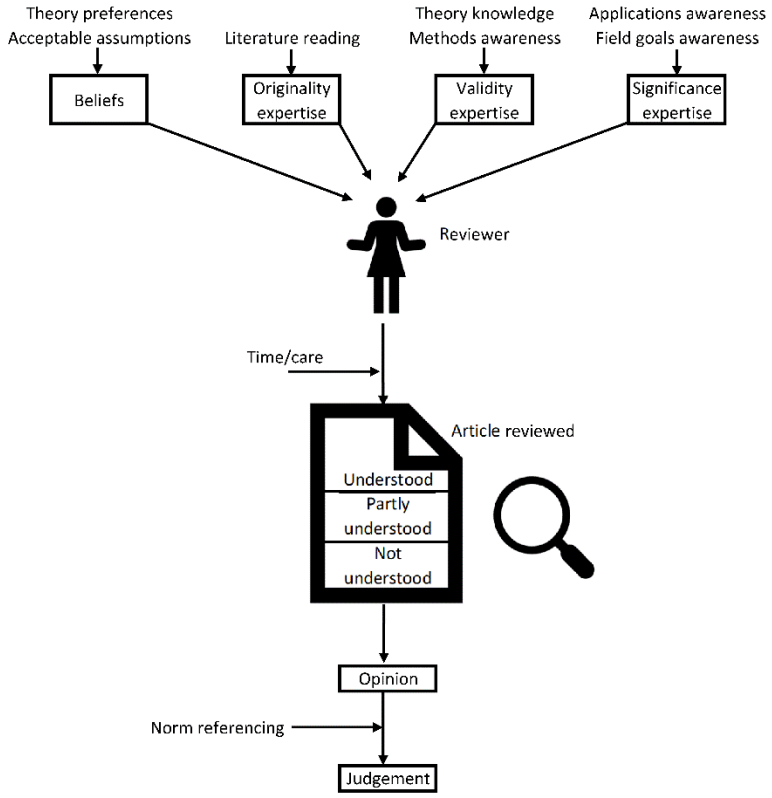

Figure 10. Model of key factors influencing peer review judgements.

Conclusion

The results of our study confirm that there is a moderate degree of agreement (ICCs 0.36 to 0.40) between journal article reviewers for each of the three core dimensions of research quality (validity, originality, and significance) in the relatively ideal case of theoretical physics journal article submissions. The reviewers agreed from 40% to 48% of the time for each facet and one-point disagreements were almost as common (39% to 46%), so large differences were rare. From a journal quality control perspective, this seems reasonable.

Despite the above conclusion, the moderate ICCs take into account that few SciPost Physics theoretical articles received scores other than 4, 5, or 6, so many reviewer differences of 0 or 1 could be expected by chance, even for reviewers assessing articles from non-overlapping perspectives. These moderate ICC values suggest that assessing the core qualities, including rigour, of a journal submission is not trivial in even contexts where this seems to be the most possible, at least in theory.

Since the results are based on a single journal, albeit for a topic that seems to be particularly likely to have a reviewer consensus, and there are many limitations, it is (weak) evidence that academia as a whole has a generic problem with reviewer consistency for originality, significance and rigour, with the last of these being the most concerning for the academic record. The disagreements may be due to differing reviewer beliefs, levels of expertise, or types of assumption that they consider acceptable. This underlines the importance of editors in selecting reliable referees that collectively can be expected to cover all relevant dimensions expertly. The results also emphasise the importance of editorial oversight to make effective judgements when reviewers almost inevitably disagree (see also: Schwartz and Zamboanga, 2009). In the absence of this, an unlucky choice of referees might allow unsound work to be published and strong work rejected. Of course, only a perfect

system would never allow errors, and this is unattainable in any human context. Thus, the scientific community might consider that journal refereeing is effective enough, with the reading audience also having the collective responsibility to critically evaluate publications and cross-check important results.

To support reviewers, it may be helpful to provide norm referencing guidelines, since this may be a major source of judgment discrepancies. Help with decisions about what are acceptable assumptions for a publishable article, if possible, would also be useful. This may even be relevant for the validity component of reviewing since few studies can provide exhaustive evidence or absolute proof. Guidelines, whilst a primary source of information for reviewers (Freda et al., 2009), are not universally consulted and can have their own problems (Sheard, 2022), so this suggestion is not a panacea.

For authors, the results suggest that it is normal to receive disagreeing reviewer reports, and this does not mean that one reviewer is bad or vindictive (such reviewers should be screened out by editors: Schwartz and Zamboanga, 2009). If a paper with disagreeing reviewers is allowed to be revised, then this gives the author the chance to modify their work to help it be regarded as high quality from a perspective that they might not have considered before.

Finally, the results also suggest that authors striving to create significant work would benefit from primarily focusing on originality rather than validity, at least in theoretical physics. This might mean spending longer on generating the initial novel idea and less time on generating exhaustive evidence to support it.

Acknowledgements. The work of MT and JAH was funded by the European Union under the Horizon Europe grant OMINO (grant number 101086321). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency. Neither the European Union nor European Research Executive Agency can be held responsible for them.

References

Aksnes, D.W., Langfeldt, L., and Wouters, P. (2019) 'Citations, citation indicators, and research quality: An overview of basic concepts and theories', *Sage Open*, 9/1: 2158244019829575.

Aksnes, D.W., Piro, F.N. and Fossum, L.W. (2023) 'Citation metrics covary with researchers' assessments of the quality of their works', *Quantitative Science Studies*. https://doi.org/10.1162/qss_a_00241

Anderson, T.N., et al. (2022) 'Surgical endoscopy education research: how are we doing?', *Surgical Endoscopy*: 1-5.

Bartko, J.J. (1966) 'The intraclass correlation coefficient as a measure of reliability', *Psychological Reports*, 19/1: 3-11.

Blunt, C. (2015) 'Hierarchies of evidence in evidence-based medicine', 'http://etheses.lse.ac.uk/3284/1/Blunt_heirachies_of_evidence.pdf

Bohannon, R.W. (1986) 'Agreement among reviewers', *Physical Therapy*, 66/9: 1431-1432.

Bonaccorsi, A. (2018) 'Peer review in social sciences and humanities. Addressing the interpretation of quality criteria. In Bonaccorsi, A. (ed.) *The evaluation of research in social sciences and humanities* (pp, 71-101) Berlin: Springer.

Bornmann, L., and Daniel, H.D. (2008) 'The effectiveness of the peer review process: Inter-referee agreement and predictive validity of manuscript refereeing at Angewandte Chemie', *Angewandte Chemie International Edition*, 47/38: 7173-7178.

Bornmann, L., and Daniel, H.D. (2010) 'Reliability of reviewers' ratings when using public peer review: a case study', *Learned Publishing*, 23/2: 124-131.

Bornmann, L., Mutz, R., and Daniel, H.D. (2010) 'A reliability-generalization study of journal peer reviews: A multilevel meta-analysis of inter-rater reliability and its determinants', *PloS One*, 5/12: e14331.

Braun, V., and Clarke, V. (2019) 'Reflecting on reflexive thematic analysis', *Qualitative Research in Sport, Exercise and Health*, 11/4: 589-597.

Capaccioni, A., and Spina, G. (2018) 'Guidelines for peer review. A survey of international practices', In *The evaluation of research in social sciences and humanities* (pp, 55-69: Berlin: Springer.

Chong, S.W., and Mason, S. (2021) 'Demystifying the process of scholarly peer-review: an autoethnographic investigation of feedback literacy of two award-winning peer reviewers', *Humanities and Social Sciences Communications*, 8/1: 1-11.

Cicchetti, D.V. (1980) 'Reliability of reviews for the American Psychologist – a biostatistical assessment of the data', *American Psychologist*, 35: 300–303.

Cicchetti, D.V. and Conn, H.O. (1978) 'Reviewer evaluation of manuscripts submitted to medical journals', *Biometrics*, 34: 728.

Cicchetti, D.V. and Eron, L.D. (1979) 'The reliability of manuscript reviewing for the Journal of Abnormal Psychology'. *Proceedings of the American Statistical Association*, 22, 596–600.

Cohen, J. (1960) 'A coefficient of agreement for nominal scales', *Educational and Psychological Measurement*, 20/1: 37-46.

Davis, W.E., et al. (2018) 'Peer-review guidelines promoting replicability and transparency in psychological science', *Advances in Methods and Practices in Psychological Science*, 1/4: 556-573.

Erosheva, E.A., Martinková, P., and Lee, C.J. (2021) 'When zero may not be zero: A cautionary note on the use of inter-rater reliability in evaluating grant peer review', *Journal of the Royal Statistical Society: Series A*, 184/3: 904-919.

Eysenbach, G. (2004) 'Improving the quality of Web surveys: the Checklist for Reporting Results of Internet E-Surveys (CHERRIES)', *Journal of Medical Internet Research*, 6/3: e132.

Falk Delgado, A., Garretson, G., and Falk Delgado, A. (2019) 'The language of peer review reports on articles published in the BMJ, 2014–2017: an observational study', *Scientometrics*, 120: 1225-1235.

Finn, R. H. (1970) 'A note on estimating the reliability of categorical data', *Educational and psychological measurement*, 30/1: 71-76.

Freda, M.C., et al. (2009) 'Peer reviewer training and editor support: results from an international survey of nursing peer reviewers', *Journal of Professional Nursing*, 25/2: 101-108.

Garcia-Costa, D., Squazzoni, F., Mehmani, B., and Grimaldo, F. (2022) 'Measuring the developmental function of peer review: A multi-dimensional, cross-disciplinary analysis of peer review reports from 740 academic journals', *PeerJ*, 10: e13539.

Guthrie, S., Ghiga, I., and Wooding, S. (2017) ‘What do we know about grant peer review in the health sciences?’ *F1000Research*, 6: 1335. https://doi.org/10.12688/f1000research.11917.2

Hamann, J., and Beljean, S. (2019) 'Academic evaluation in higher education’. In: Teixeira, P. (ed.) International Encyclopedia of Higher Education Systems and Institutions. Berlin: Springer (pp 28-34).

Hargens, L., and Herting, J. (1990a) 'Neglected considerations in the analysis of agreement among journal referees', *Scientometrics*, 19/1-2: 91-106.

Hargens, L., and Herting, J. (1990b) 'A new approach to referees' assessments of manuscripts', *Social Science Research*, 19/1: 1-16.

Ho, R.C.M., et al. (2013) ‘Views on the peer review system of biomedical journals: an online survey of academics from high-ranking universities’, *BMC Medical Research Methodology*, 13: 1-15.

Horbach, S. P., Tijdink, J. K., and Bouter, L. M. (2022) ‘Partial lottery can make grant allocation more fair, more efficient, and more diverse’, Science and Public Policy, 49/4: 580-582.

Hug, S. E. (2022) 'Towards theorizing peer review', *Quantitative Science Studies*. https://doi.org/10.1162/qss_a_00195

Jadad, A. R., et al. (1996) 'Assessing the quality of reports of randomized clinical trials: is blinding necessary?', *Controlled Clinical Trials*, 17/1: 1-12.

Jerrim, J., and Vries, R. D. (2020) ‘Are peer-reviews of grant proposals reliable? An analysis of Economic and Social Research Council (ESRC) funding applications’, *The Social Science Journal*, 60/1: 91-109.

Jirschitzka, J., Oeberst, A., Göllner, R., and Cress, U. (2017) 'Inter-rater reliability and validity of peer reviews in an interdisciplinary field', *Scientometrics*, 113/2: 1059-1092.

Key, J., et al. (2006) 'Meta-analysis of studies of alcohol and breast cancer with consideration of the methodological issues', *Cancer Causes and Control*, 17/6: 759-770.

Kirk, S.A., and Franke, T.M. (1997) 'Agreeing to disagree: A study of the reliability of manuscript reviews', *Social Work Research*, 21/2: 121–126. https://doi.org/10.1093/swr/21.2.121.

Kitchenham, B. A., et al. (2012) 'Three empirical studies on the agreement of reviewers about the quality of software engineering experiments', *Information and Software Technology*, 54/8: 804-819.

Koo, T.K., and Li, M.Y. (2016) 'A guideline of selecting and reporting intraclass correlation coefficients for reliability research', *Journal of Chiropractic Medicine*, 15/2: 155-163.

Langfeldt, L., Nedeva, M., Sörlin, S., and Thomas, D.A. (2020) 'Co-existing notions of research quality: A framework to study context-specific understandings of good research', *Minerva*, 58/1: 115-137.

Lee, C.J., Sugimoto, C.R., Zhang, G., and Cronin, B. (2013) ‘Bias in peer review’, *Journal of the American Society for Information Science and Technology*, 64/1: 2-17.

Liljequist, D., Elfving, B., and Skavberg Roaldsen, K. (2019) ‘Intraclass correlation–A discussion and demonstration of basic features’, PloS One, 14/7, e0219854.

Linden, W., Craig, K.D., and Wen, F.K. (1992) 'Contributions of reviewer judgements to editorial decision-making for the Canadian Journal of Behavioural Science: 1985–1986', *Canadian Journal of Behavioural Science*, 24/4: 433.

Maggin, D.M., Chafouleas, S.M., Berggren, M., and Sugai, G. (2013) 'A systematic appraisal of peer review guidelines for special education journals', *Exceptionality*, 21/2: 87-102.

Marson, S.M., and Lillis, J.P. (2022) 'A case study for the interrater reliability of journal referees', *Research on Social Work Practice*, 32/2: 238-244.

Marušić, A., et al. (2002) 'Peer review in a small and a big medical journal: case study of the Croatian Medical Journal and the Lancet', *Croatian Medical Journal*, 43/3: 286-289.

Minelli, E., Rebora, G., and Turri, M. (2008) 'The structure and significance of the Italian research assessment exercise (VTR)', In *European Universities in Transition: Issues, models and cases*, (pp. 221-236).

Morrow, J.R., Bray, M.S., Fulton, J.E., and Thomas, J.R. (1992) 'Interrater reliability of 1987–1991 Research Quarterly for Exercise and Sport reviews', *Research Quarterly for Exercise and Sport*, 63/2: 200-204.

Munley, P.H., Sharkin, B.S., and Gelso, C.J. (1988) 'Reviewer ratings and agreement on manuscripts reviewed for the Journal of Counseling Psychology', *Journal of Counseling Psychology*, 35/2: 198-202.

Neuendorf, K.A. (2017) *The content analysis guidebook (2 ed),* Oxford: Sage.

Oxman, A.D., et al. (1991) 'Agreement among reviewers of review articles', *Journal of Clinical Epidemiology*, 44/1: 91-98.

Peters, D.P., and Ceci, S.J. (1982) 'Peer-review practices of psychological journals: The fate of published articles, submitted again', *Behavioral and Brain Sciences*, 5/2: 187-195.

Peterson, D.A. (2020) 'Dear reviewer 2: Go f' yourself', *Social Science Quarterly*, 101/4: 1648-1652.

Plug, C. (1993) 'The reliability of manuscript evaluation for the South African Journal of Psychology', *South African Journal of Psychology*, 23/1: 43-48.

Ritson, S. (2021) 'Constraints and divergent assessments of fertility in non-empirical physics in the history of the string theory controversy', *Studies in History and Philosophy of Science Part A*, 90: 39-49.

Rothwell, P.M., and Martyn, C.N. (2000) 'Reproducibility of peer review in clinical neuroscience: Is agreement between reviewers any greater than would be expected by chance alone?', *Brain*, 123/9: 1964-1969.

Schroter, S., Tite, L., Hutchings, A., and Black, N. (2006) 'Differences in review quality and recommendations for publication between peer reviewers suggested by authors or by editors', *Jama*, 295/3: 314-317.

Schwartz, S.J., and Zamboanga, B.L. (2009) 'The peer-review and editorial system: Ways to fix something that might be broken', *Perspectives on Psychological Science*, 4/1: 54-61.

Scott, W.A. (1974) 'Interreferee agreement on some characteristics of manuscripts submitted to the Journal of Personality and Social Psychology', *American Psychologist*, 29/9: 698.

Seeber, M., et al. (2021) 'Does reviewing experience reduce disagreement in proposals evaluation? Insights from Marie Skłodowska-Curie and COST Actions', *Research Evaluation*, 30/3: 349-360.

Seeber, M. (2020) 'How do journals of different rank instruct peer reviewers? Reviewer guidelines in the field of management', *Scientometrics*, 122/3: 1387-1405.

Sheard, L. (2022) 'Telling a story or reporting the facts? Interpretation and description in the qualitative analysis of applied health research data: A documentary analysis of peer review reports', *SSM-Qualitative Research in Health*, 2: 100166.

Shepherd, C., and Challenger, R. (2013) 'Revisiting paradigm (s) in management research: A rhetorical analysis of the paradigm wars', *International Journal of Management Reviews*, 15/2: 225-244.

Shrout, P.E., and Fleiss, J.L. (1979) 'Intraclass correlations: uses in assessing rater reliability', *Psychological Bulletin*, 86/2: 420.

Song, E., et al. (2021) 'A scoping review on biomedical journal peer review guides for reviewers', *PloS One*, 16/5: e0251440.

Sposato, L.A., et al. (2014) 'A peek behind the curtain: Peer review and editorial decision making at Stroke', *Annals of Neurology*, 76/2: 151-158.

Strevens, M. (2021). *The knowledge machine: How irrationality created modern science*, New York: Liveright Publishing.

Superchi, C., et al. (2020) 'Development of ARCADIA: a tool for assessing the quality of peer-review reports in biomedical research', *BMJ Open*, 10/6: e035604.

Tennant, J.P., and Ross-Hellauer, T. (2020) 'The limitations to our understanding of peer review', *Research Integrity and Peer Review*, 5/1: 1-14.

Thelwall, M. (2022) 'Journal and disciplinary variations in academic open peer review anonymity, outcomes, and length', *Journal of Librarianship and Information Science*, https://doi.org/10.1177/09610006221079345

Thelwall, M., et al. (2021) 'Does the use of open, non-anonymous peer review in scholarly publishing introduce bias? Evidence from the F1000Research post-publication open peer review publishing model', *Journal of information science*, 47/6: 809-820.

Tinsley, H.E., and Weiss, D.J. (1975) 'Interrater reliability and agreement of subjective judgments', *Journal of Counseling Psychology*, 22/4: 358.

Tourish, D. (2020) 'The triumph of nonsense in management studies', *Academy of Management Learning and Education*, 19/1: 99-109.

Traag, V.A., Malgarini, M., and Sarlo, S. (2020) 'Metrics and peer review agreement at the institutional level. arXiv preprint arXiv:2006',14830.

Travis, G.D.L., and Collins, H.M. (1991) 'New light on old boys: Cognitive and institutional particularism in the peer review system', *Science, Technology, & Human Values*, 16/3: 322-341.

Van Rooyen, S., Black, N., and Godlee, F. (1999) 'Development of the review quality instrument (RQI) for assessing peer reviews of manuscripts', *Journal of Clinical Epidemiology*, 52/7: 625-629.

Wade, D., and Tennant, A. (2004) 'An audit of the editorial process and peer review in the journal Clinical rehabilitation', Clinical Rehabilitation, 18/2: 117-124.

Warne, V. (2016) 'Rewarding reviewers–sense or sensibility? A Wiley study explained', *Learned Publishing*, 29/1: 41-50.

Whitehurst, G.J. (1983) 'Interrater agreement for reviews for Developmental Review', *Developmental Review*, 3/1: 73-78.

Whitley, R. (2000) *The intellectual and social organization of the sciences*, Oxford: Oxford University Press.

Wolfram, D., Wang, P., Hembree, A., and Park, H. (2020) 'Open peer review: promoting transparency in open science', *Scientometrics*, 125/2: 1033-1051.